DAT405 Introduction to Data Science and AI

Assignment 4: Spam classification using Naïve Bayes

Francisco Boudagh (13 h)

Jakob Engström (13 h)

2023-04-26

The three folders containing emails are downloaded manually.

1. Preprocessing

1.1 Look at a few emails from easy_ham, hard_ham and spam. Do you think you would be able to classify the emails just by inspection? How do you think a succesful model can learn the difference between the different classes of emails?

In [23]:
```
import email.policy
from email import message_from_string
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
import os

def extract_emails(file, class_name):
  emails_list = []
  for directory in file:
    for file in os.listdir(directory):
      with open(os.path.join(directory,file), encoding='latin-1') as f:
        if f:
          content = f.read()
          emails_list.append({'email': message_from_string(content, policy=email.policy.default), 'content': content, 'class': class_name})
  return pd.DataFrame(emails_list)

# extract emails and add class information
df_eh = extract_emails(['easy_ham'], 'eham')
df_hh = extract_emails(['hard_ham'], 'hham')
df_s = extract_emails(['spam'], 'spam')

# combining easy ham, hard ham and spam
df_combined = pd.concat([df_eh, df_hh, df_s])

print('Easy ham:', + len(df_eh))
print('Hard ham:', + len(df_hh))
print('Spam:', + len(df_s))

pd.set_option('max_colwidth',4500)

# printing some spam samples to find patterns manually
df_s.sample(5)
```

```
Easy ham: 2551
Hard ham: 250
Spam: 501
```

Out[23]:

	email	
93	[Return-Path, Delivered-To, Received, Received, Received, Message-Id, From, To, Cc, Subject, Sender, MIME-Version, Date, X-Mailer, Content-Type]	From tawneemail@yahoo.com Mon Aug 26 15:49:20 2002\nReturn-Path: <tawneemail@yahoo.com>\nDelivered-To: zzzz@localhost.example.com\nReceived: from localhost (localhost [127.0.0.1])\n\tby phobos.labs.example.com (Postfix) with ESMTP id D96FE47CC8\n\tfor <zzzz@localhost>; Mon, 26 Aug 2 (single-drop); Mon, 26 Aug 2002 15:41:57 +0100 (IST)\nReceived: from proxy.foxberry.com ([203.238.133.122]) by\n dogma.slashnull.org (8.11.6/8.11.6) with SMTP id g7PBWaZ11279 for\n <webmaster@efi.ie>; Sun, 25 Aug 2002 12:32:37 +0100\nMessage-Id: <200208251132.g7PBWaZ11279@dogma.slashnul ready to bare for you!\nSender: Tawnee Stone <tawneemail@yahoo.com>\nMIME-Version: 1.0\nDate: Sun, 25 Aug 2002 19:33:54 -0700\nX-Mailer: Internet Mail Service (5.5.2650.21)\nContent-Type: text/html; charset="iso-8859-1"\n\n<TABLE borderColorDark=#0000cc width=700 align=center bgColor=#f Sensation is ready to bare all for \nyou.</B> <B>Looking at those <FONT color=#0000ff>firm tits and creamy soft \nthighs</FONT> you can't help but get an instant erection. Now Tawnee's Sister \nTori joined in. Both are sweet little honeys that love to show you what a color="#FF0000">TawneeStone.com</A></FONT><FONT color=#000000> \n where you will see both posing for you in many erotic ways</FONT></B> \n </H3>\n </DIV></TD></TR></TBODY></TABLE>\n\t <TABLE borderColorDark=#0000cc width=700 align=center bgColor=#ffff Helvetica, sans-serif">THIS \n EMAIL WAS NOT SENT UNSOLICITED. YOU HAVE SUBSCRIBED TO RECEIVE OUR MAILING. \n IF YOU WISH TO DISCONTINUE SIMPLY REPLY TO THIS EMAIL WITH "REMOVE
274	[Return-Path, Delivered-To, Received, Received, Received, Received, Received, Message-Id, Date, From, Reply-To, To, Subject, MIME-Version, X-Ema-Cid, X-Ema-Lid, X-Ema-PC, Content-Type]	From bounce@trafficmagnet.com Tue Sep 10 11:13:55 2002\nReturn-Path: <bounce@trafficmagnet.com>\nDelivered-To: zzzz@localhost.jmason.org\nReceived: from localhost (jalapeno [127.0.0.1])\n\tby zzzzason.org (Postfix) with ESMTP id B09FC16F03\n\tfor <zzzz@localhost>; Tue, 10 Sep 2002 11:13:53 + Tue, 10 Sep 2002 11:13:53 +0100 (IST)\nReceived: from webnote.net (mail.webnote.net [193.120.211.219]) by\n dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g89JZ0C29435 for\n <zzzz-list-admin-iiu@jmason.org>; Mon, 9 Sep 2002 20:35:05 +0100\nReceived: from localhost.localdomain ([211.101.236.18 +0100\nReceived: from () by localhost.localdomain (8.11.6/8.11.6) with ESMTP id\n g89JRj329191 for <zzzz-list-admin-iiu@jmason.org>; Tue, 10 Sep 2002 03:27:51\n +0800\nMessage-Id: <505CR1000010979@p1j2m3a4.pdhost.com>\nDate: Tue, 10 Sep 2002 03:34:01 +0800 (CST)\nFrom: Sarah iiu@jmason.org\nSubject: iiu.taint.org\nMIME-Version: 1.0\nX-Ema-Cid: 11513610\nX-Ema-Lid: \nX-Ema-PC: 0f02f9b386000\nContent-Type: multipart/alternative; boundary=1189366250.1031600041421.JavaMail.SYSTEM.emaserver2\n\n--1189366250.1031600041421.JavaMail.SYSTEM.emaserver2\nConten on some search engines! I think we can\noffer you a service which can help you increase traffic and the number of visitors to your website.\n\nI would like to introduce you to TrafficMagnet.com. We offer a unique technology that will submit your \nwebsite to over 300,000 search engines and directories every mont TrafficMagnet and the cost for submitting your website to over 300,000 search \nengines and directories, visit us at:\n\nhttp://p1j2m3a4.pdhost.com/pdsvr/www/r?1000010979.505.23.eUcTDXPhW5zo\n \n\nI would love to hear from you. \n\n\nBest Regards,\n\nSarah Williams\nSales and Marketing \nE-mail: S understand that you may NOT wish to receive information from me by email. \nTo be removed from this and other offers, simply go to the link below:\nhttp://p1j2m3a4.pdhost.com/pdsvr/www/optoutredirect?UC=Lead&UI=11513610\n--1189366250.1031600041421.JavaMail.SYSTEM.emaserver2\nContent-Ty TYPE="text/css">\n<!--\nTD { font-family: verdana, arial, helvetica; font-size: 11px; color: #000000 }\n-->\n</STYLE>\n</HEAD>\n<BODY BGCOLOR="#FFFFFF">\n<TABLE WIDTH="600" BORDER="0" CELLSPACING="0" CELLPADDING="0">\n<TR>\n\t<TD>Hi \n\t \n\tI visited <A HREF="http://p1j some search engines! I think we can offer \n\tyou a service which can help you increase traffic and the number of visitors \n\tto your website. \n\t \n\tI would like to introduce you to <A HREF="http://p1j2m3a4.pdhost.com/pdsvr/www/r?1000010979.505.2.Wcdniklsi+C9l6">TrafficMagnet.com</A> \n\t \n\t<TABLE WIDTH="398" BORDER="0" CELLSPACING="0" CELLPADDING="0" ALIGN="center">\n\t<TR>\n\t\t<TD><A HREF="http://p1j2m3a4.pdhost.com/pdsvr/www/r?1000010979.505.8.43CuFchAOTr0t7"><IMG SRC="http://www.trafficmagnet.com/img/img_tm.gif" 1000010979.505.11.5J2PcvS63vRqi$"><IMG SRC="http://www.trafficmagnet.com/img/img_website.gif" WIDTH="197" HEIGHT="141" BORDER="1"></A></TD>\n\t\t<TD VALIGN="bottom"><A HREF="http://p1j2m3a4.pdhost.com/pdsvr/www/r?1000010979.505.14.dbH9rBk43Z3NzY"><IMG SRC="http:/
71	[Return-Path, Delivered-To, Received, Received, Received, From, Reply-To, Message-ID, To, Subject, Date, MiME-Version, Content-Type, X-Priority, X-MSMail-Priority, X-Mailer, Importance, Content-Transfer-Encoding]	From claudia_robinson@eudoramail.com Mon Aug 26 15:15:14 2002\nReturn-Path: <claudia_robinson@eudoramail.com>\nDelivered-To: zzzz@localhost.example.com\nReceived: from localhost (localhost [127.0.0.1])\n\tby phobos.labs.example.com (Postfix) with ESMTP id 722F043F9B\n\tfor <zzzz@local (fetchmail-5.9.0)\n\tfor zzzz@localhost (single-drop); Mon, 26 Aug 2002 15:13:33 +0100 (IST)\nReceived: from eudoramail.com (c17996.rivrw4.nsw.optusnet.com.au [211.28.162.97])\n\tby webnote.net (8.9.3/8.9.3) with SMTP id UAA22962\n\tfor <zzzz@example.com>; Sun, 25 Aug 2002 20:36:12 +0100\nFrom: 3db10ed5@qclywl>\nTo: <zzzz@example.com>\nSubject: re: domain registration savings\nDate: Mon, 26 Aug 2002 06:30:21 -1100\nMiME-Version: 1.0\nContent-Type: text/plain;\n\tcharset="iso-8859-1"\nX-Priority: 3 (Normal)\nX-MSMail-Priority: Normal\nX-Mailer: Internet Mail Service (5.5.2650.21)\nImportance the general public at discount prices. 14.95.ThesebrandnewdomainextensionswererecentlyapprovedbyICANNandhavethesamerightsastheoriginal.COMand.NETdomainnames.Thebiggestbenefitisof−coursethatthe.BIZand.INFOdomainnamesarecurrentlymoreavailable.i.e.itwillbemucheasier .affordable−domains.comtodayformoreinfo.\n\nRegisteryourdomainnametodayforjust 14.95 at: http://www.affordable-domains.com/ Registration fees include full access to an easy-to-use control panel to manage your domain name in the future.\n \nSincerely,\n \nDomain Administrator\nAffordable Domains\n\n\nTo remove your email address from further promotional mailings from this co
352	[Return-Path, Delivered-To, Received, Received, Received, Message-Id, Errors-To, From, Organisation, To, Subject, Date, MIME-Version, Content-Type]	From AdvertSpRj40@cs.com Tue Sep 17 11:41:33 2002\nReturn-Path: <AdvertSpRj40@cs.com>\nDelivered-To: zzzz@localhost.jmason.org\nReceived: from localhost (jalapeno [127.0.0.1])\n\tby zzzzason.org (Postfix) with ESMTP id 0948E16F03\n\tfor <zzzz@localhost>; Tue, 17 Sep 2002 11:40:46 +0100 (IS Sep 2002 11:40:46 +0100 (IST)\nReceived: from cs.com (12-245-142-206.client.attbi.com [12.245.142.206])\n by dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g8H9cKC04150 for\n <zzzz@jmason.org>; Tue, 17 Sep 2002 10:38:21 +0100\nMessage-Id: <4174-22002921794043650@cs.com>\nErrors-To: Ad & SERVICES\nTo: "zzzz@jmason.org" <jm@jmason.org>\nSubject: May I have a moment of your Time PLEASE\nDate: Tue, 17 Sep 2002 05:40:44 -0400\nMIME-Version: 1.0\nContent-Type: multipart/mixed; boundary="----=_NextPart_000_01BC2B74.89D1CCC0"\n\nThis is a multi-part me =_NextPart_84815C5ABAF209EF376268C8"\n\n------=_NextPart_84815C5ABAF209EF376268C8\nContent-type: text/plain; charset=windows-1252\nContent-Transfer-Encoding: quoted-printable\n\nDear Sir/Madam\n\nWishing you a wonderful day=2E With an offer, to save you money and time=\n=2E Shop store=\n offers over 1500 quality products, with 15 plus categories=2E\nFor you to, window shop through=2E=20\n\nThe following are E-coupons numbers, which you can use with any order=2E Y=\nou place from Johnson Home Products Online Super Store=2E\nDuring the check out process of our online secure on any order you place with J=\nohnson Home Products Online Store=2E All major Credit cards\naccepted=2E\n\nPlease visit us at!=20\nhttp://www=2Ejohnsonhome2276=2Ecom\n\nE-Coupon Numbers:\n\nMJ95L594568JWWL 5% off any order ! E-coupon Expires's 01/14/2003 =\nany orders, maximum cash email list you may have sub=\nscript to or from the search engine emailing list, from your submission to=\n search engines=2E If you have received this email by error, I am sorry fo=\nr any inconvenience=2E To be remove from the emailing list, reply to this =\nemail with the word remove in the subject area of th with mail in orders=2E\n\nThank you for your time!\n\nGod Bless you!\nSincerely,\n\nBob Johnson\nEmail me with any concerns, comments, or problems, or to be removed from t=\nhe emailing list=2E\nAdvertSpRj40@cs=2Ecom\n\nTo be removed from my Email list Please reply to this email=2E type "re charset=windows-1252\nContent-Transfer-Encoding: quoted-printable\n\n<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4=2E0 Transitional//EN">\n<html>\n<head>\n<title>Welcome</title>\n</head>\n<body>\n<center>\n< color=3D"#6633ff"><b> Click=\n on to continue \n</center>\n\n</body>\n</html>\n\n------=_NextPart_84815C5ABAF209EF376268C8--\n\n------=_NextPart_000_01BC2B74.89D1CCC0\nContent-Type: image/gif; filename="ecp10cc.gif"\n\nR0lGODlhfQIGAfcAAPn6/er3/PPp/Nv3+/b47v0CA+no+vTo7Nrm7Nvn+AEB2jEzAtn07gIB4trY\n9+jZ+fHa+wEjAuj17fDL+fPY6eTJ6+bK+ObY68z3+8rX9wALAMvm7Mvn+MnI9veJjfpDPce79ebY\n3LGy8flJTPawtPR3iPO1x/eEfdnY68jX6v0aJLbN8Ofm7PdscPovMnyE8fs7Rerk3NnJ94yN7trY
165	[Return-Path, Delivered-To, Received, Received, Received, Message-Id, From, Reply-To, To, Date, Subject, X-Priority, X-Mailer, MIME-Version, Content-Type, X-MIME-Autoconverted, Content-Transfer-Encoding]	From IKE_EJOH@YAHOO.COM Thu Aug 29 15:40:40 2002\nReturn-Path: <IKE_EJOH@YAHOO.COM>\nDelivered-To: zzzz@localhost.example.com\nReceived: from localhost (localhost [127.0.0.1])\n\tby phobos.labs.example.com (Postfix) with ESMTP id 12B5643F99\n\tfor <zzzz@localhost>; Thu, 29 Aug 20 zzzz@localhost (single-drop); Thu, 29 Aug 2002 15:40:40 +0100 (IST)\nReceived: from 2mails71.com ([217.78.76.157])\n\tby webnote.net (8.9.3/8.9.3) with SMTP id PAA10752\n\tfor <zzzz@example.com>; Thu, 29 Aug 2002 15:33:45 +0100\nMessage-Id: <200208291433.PAA10752@webnote.net>\nFrom: "MR 14:41:06 +0200\nSubject: YOUR ANTICIPATED ASSISTANT IS REQUIRED.\nX-Priority: 1\nX-Mailer: Microsoft Outlook Express 5.00.2919.6900 DM\nMIME-Version: 1.0\nContent-Type: text/plain; charset="us-ascii"\nX-MIME-Autoconverted: from quoted-printable to 8bit by webnote.net id PAA10752\nContent-T confidential business proposal for you. On June 6 1999 a FOREIGN Oil consultant/contractor with the FEDE 25,000,000.00(Twenty−fiveMillionDollars)inmybranch.Uponmaturity,Isentaroutinenotificationtohisforwardingaddressbutgotnoreply.Afteramonth,wesentareminderandfinallywediscoveredfromhisemployers,theFEDERALMINISTRYOFAVIATIONtha IfoundoutthathediedwithoutmakingaWILL,andallattemptstotracehisnextofkinwasfruitless.IthereforemadefurtherinvestigationanddiscoveredthatMr.BarryKellydidnotdeclareanykinorrelationsinallhisofficialdocuments,includinghisBankDepositpaperworki 25,000,000.00 is still sitting in my Bank and the interest is being rolled over with the principal sum at the end of each year. No one willever come forward to claim it.According to Nigerian Law, at the expiration of 6 (six) years, the money will revert to the ownership of the Nigerian Government if nobody applies to cl BASIC FINANCIAL COMPANY., I will like you to provide immediately your full names and address so that i will prepare the necessary documents and affidavits, which will put you in place as the owner of this fund in the security company. i shall employ the service of an Attorney for drafting and notarization of the c the paperwork for this transaction will be done by the Attorney and my position as the Bran\nch Manager guarantees the successful execution of this transaction. If you are interested, please replyimmediately via the private email address . Upon your response, I shall then provide you with more details and releva transaction would be most profitable for both of us because I shall require your assistance

Answer 1.1:

We can see that spam mails fearue a lot of cursive and bold letters, mainly in black and red. A reocurring theme is money. So yes, some of the mails are quite easy for us to classify. If given a big amount of test data a model could probably diferentiate spam from ham quite easily. Som of the indicators could be as written before, the color and type of text. Another indicator could be to look after some reoucurring words and phrase such as 'money', 'bitcoin' and such. Also looking at the content and determining wether all the words are in a common dictionary or not since random letter combinations are reocurring in these spam mails.

1.2 Note that the email files contain a lot of extra information, besides the actual message. Ignore that for now and run on the entire text (in the optional part further down can experiment with filtering out the headers and footers). We don't want to train and test on the same data (it might help to reflect on why if you don't recall). Split the spam and the ham datasets in a training set and a test set. (`hamtrain` , `spamtrain` , `hamtest` , and `spamtest`). Use only the easy_ham part as ham data for quesions 1 and 2.

In [19]:
```
from sklearn.model_selection import train_test_split

df_eh = extract_emails(['easy_ham'], 'eham')
df_s = extract_emails(['spam'], 'spam')

# join the two dataframes only for easy_ham and spam as instructions
df_combined_eh_s = pd.concat([df_eh, df_s])

X = df_combined_eh_s['content']
y = df_combined_eh_s['class']

# splitting training data, train size: 65%, test size: 35%
hamtrain, hamtest, spamtrain, spamtest = train_test_split(X, y, train_size=0.65)

# create new dataframes for each set'
hamtrain = pd.DataFrame({'content': hamtrain, 'class': hamtrain})
hamtest = pd.DataFrame({'content': hamtest, 'class': hamtest})
spamtrain = pd.DataFrame({'content': spamtrain, 'class': spamtrain})
spamtest = pd.DataFrame({'content': spamtest, 'class': spamtest})
```

2.1 Write a Python program that:

1. Uses the four datasets from Question 1 (`hamtrain` , `spamtrain` , `hamtest` , and `spamtest`)

2. Trains a Naïve Bayes classifier (use the scikit-learn library) on `hamtrain` and `spamtrain` , that classifies the test sets and reports True Positive and False Negative rates on the `hamtest` and `spamtest` datasets. Use `CountVectorizer` (Documentation here) to transform the email texts into vectors. Please note that there are different types of Naïve Bayes Classifier in scikit-learn (Documentation here). Test two of these classifiers that are well suited for this problem:

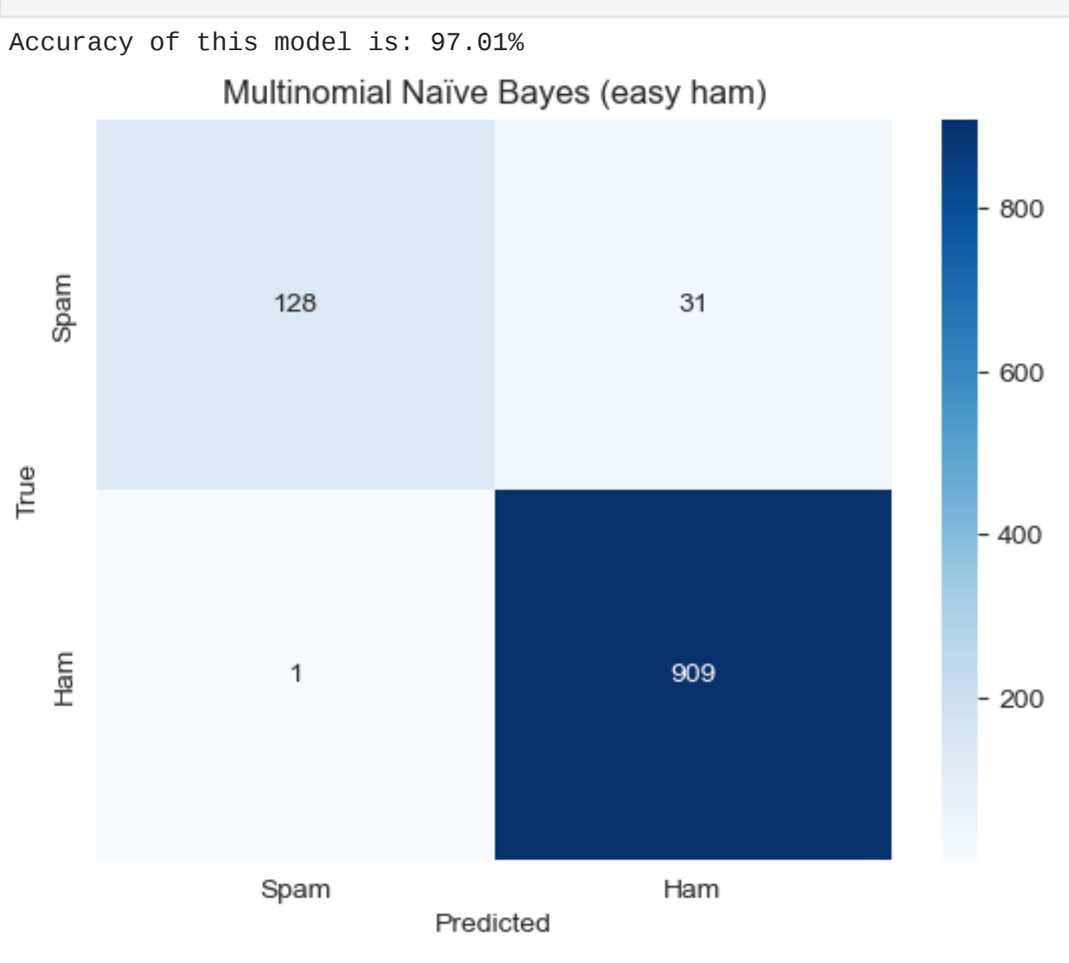
- Multinomial Naive Bayes

- Bernoulli Naive Bayes.

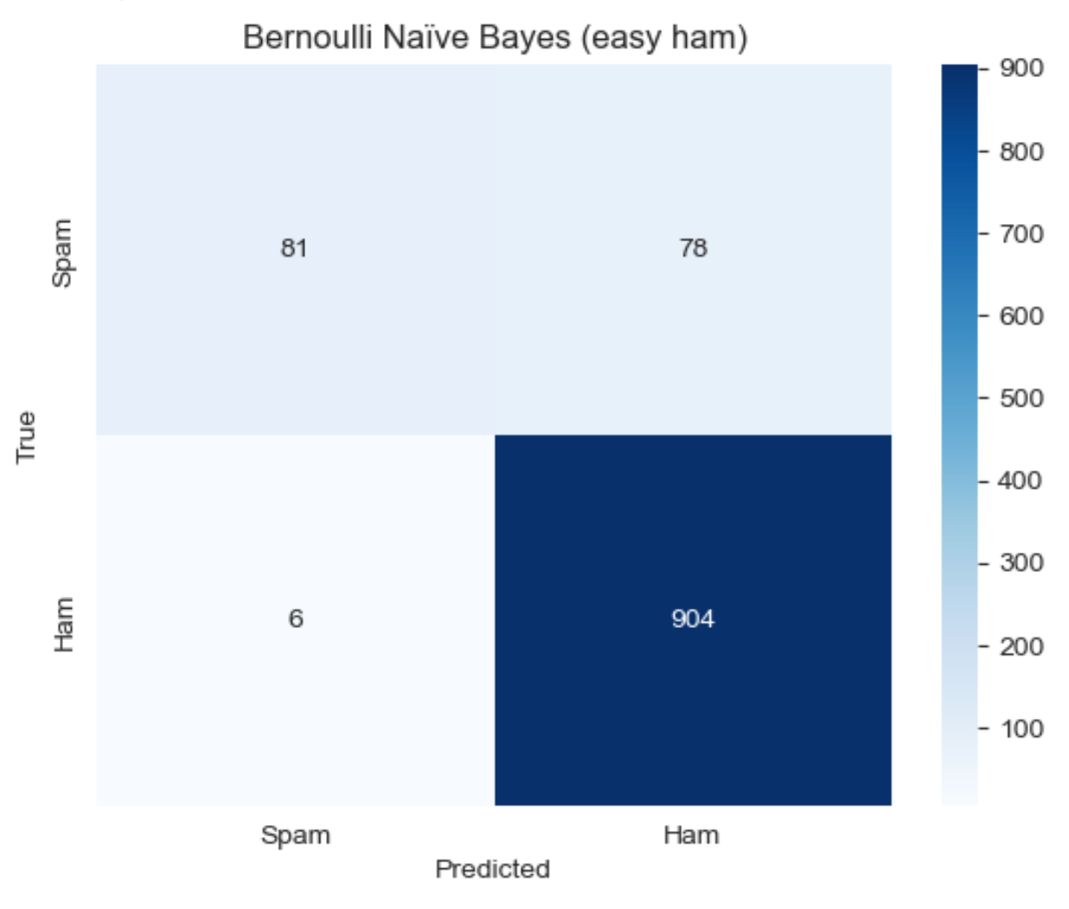
In [11]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import confusion_matrix
from sklearn.naive_bayes import BernoulliNB
from sklearn.metrics import accuracy_score

# cm labels
categories= ["Spam", "Ham"]
group_names = ['True ham','False spam','False ham','True spam']

def spam_detection_cm(hamtrain, hamtest, spamtrain, spamtest, categories, groupnames, cm_title, classifier):
    vectorizer = CountVectorizer()
    hamtrain = vectorizer.fit_transform(hamtrain)
    hamtest = vectorizer.transform(hamtest)

    if classifier == 'MNB':
        model = MultinomialNB()
    elif classifier == 'BNB':
        model = BernoulliNB()
    else:
        raise ValueError("Invalid classifier type. Please choose either 'MNB' or 'BNB'.")

    # fit train data into the chosen model
    model.fit(hamtrain, spamtrain)

    # accuracy score in %
    accuracy = model.score(hamtest, spamtest) * 100
    print("Accuracy of this model is: {:.2f}%".format(accuracy))

    spam_test_pred = model.predict(hamtest)
    cm = confusion_matrix(spamtest, spam_test_pred)

    group_counts = ["{0:0.0f}".format(value) for value in cm.flatten()]
    group_percentages = ["{0:.2%}".format(value) for value in cm.flatten()/np.sum(cm)]
    df_cm = pd.DataFrame(cm, range(2), range(2))

    sns.set_style("whitegrid", {'axes.grid' : False})
    ax = sns.heatmap(df_cm, annot=True, fmt='', cmap='Blues', xticklabels=categories, yticklabels=categories)
    ax.set_title(cm_title)
    ax.set_ylabel("True")
    ax.set_xlabel("Predicted")
    plt.show()
```

In [12]:
```
df_ham_spam = pd.concat([df_eh, df_s])

# spam = 1 and ham = 0
df_ham_spam.loc[df_ham_spam['class'] == 'spam', 'class'] = 1
df_ham_spam.loc[df_ham_spam['class'] == 'ham', 'class'] = 0

# create datasets with 65& train size and 35% test size
hamtrain, hamtest, spamtrain, spamtest = train_test_split(df_ham_spam['content'], df_ham_spam['class'].values.tolist(), train_size=0.65)

spam_detection_cm(hamtrain, hamtest, spamtrain, spamtest, categories, group_names, "Multinomial Naïve Bayes (easy ham)", 'MNB')
spam_detection_cm(hamtrain, hamtest, spamtrain, spamtest, categories, group_names, "Bernoulli Naïve Bayes (easy ham)", 'BNB')
```

```
Accuracy of this model is: 97.01%
```



```
Accuracy of this model is: 92.14%
```



2.2 Answer the following questions:

a) What does the CountVectorizer do?

The function CountVectorizer transforms text into numerical matrix by first splitting the words and then counting their frequency. This is since we need numerical values inorder to use the classsfiers.

b) What is the difference between Multinomial Naive Bayes and Bernoulli Naive Bayes

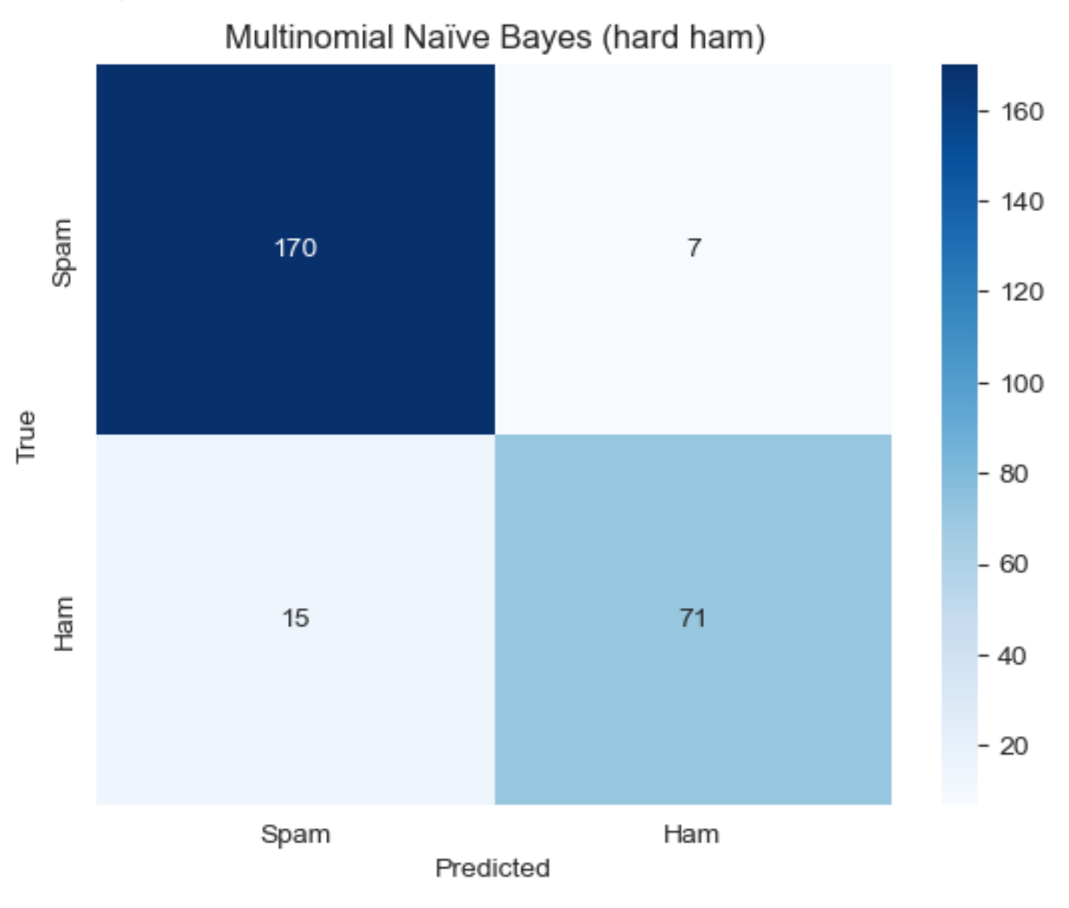
Bernoulli Naive Bayes is a binary model. It only sees if a word (from trained dataset) exists (in test mail) or not (so it will give a value, 0 or 1), while Multinomial count the reoucurrance of each word, which normalized would be a value from 0 to 1. So it is reasonable (as the result above show) that Multinomial would perform better in this case since the data represents words.

3.1 Run the two models:

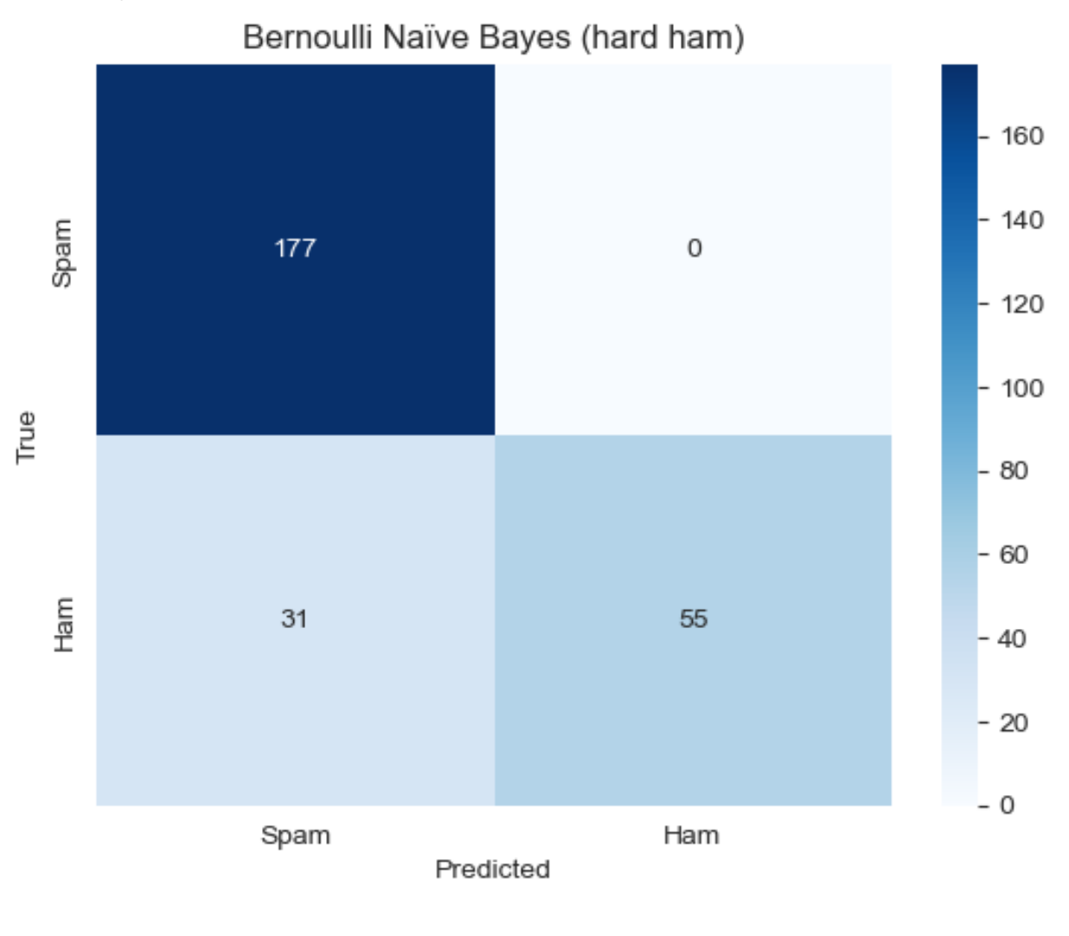
Run (don't retrain) the two models from Question 2 on spam versus hard-ham. Does the performance differ compared to question 2 when the model was run on spam versus easy-ham? If so, why?

In [13]:
```
df_ham_spam = pd.concat([df_hh, df_s])

# spam = 1 and ham = 0
df_ham_spam.loc[df_ham_spam['class'] == 'spam', 'class'] = 1
df_ham_spam.loc[df_ham_spam['class'] == 'ham', 'class'] = 0

# Create datasets with 65& train size and 35% test size
hamtrain, hamtest, spamtrain, spamtest = train_test_split(df_ham_spam['content'], df_ham_spam['class'].values.tolist(), train_size=0.65)

spam_detection_cm(hamtrain, hamtest, spamtrain, spamtest, categories, group_names, "Multinomial Naïve Bayes (hard ham)", 'MNB')
spam_detection_cm(hamtrain, hamtest, spamtrain, spamtest, categories, group_names, "Bernoulli Naïve Bayes (hard ham)", 'BNB')
```

```
Accuracy of this model is: 91.63%
```



```
Accuracy of this model is: 88.21%
```



Answer 3.1:

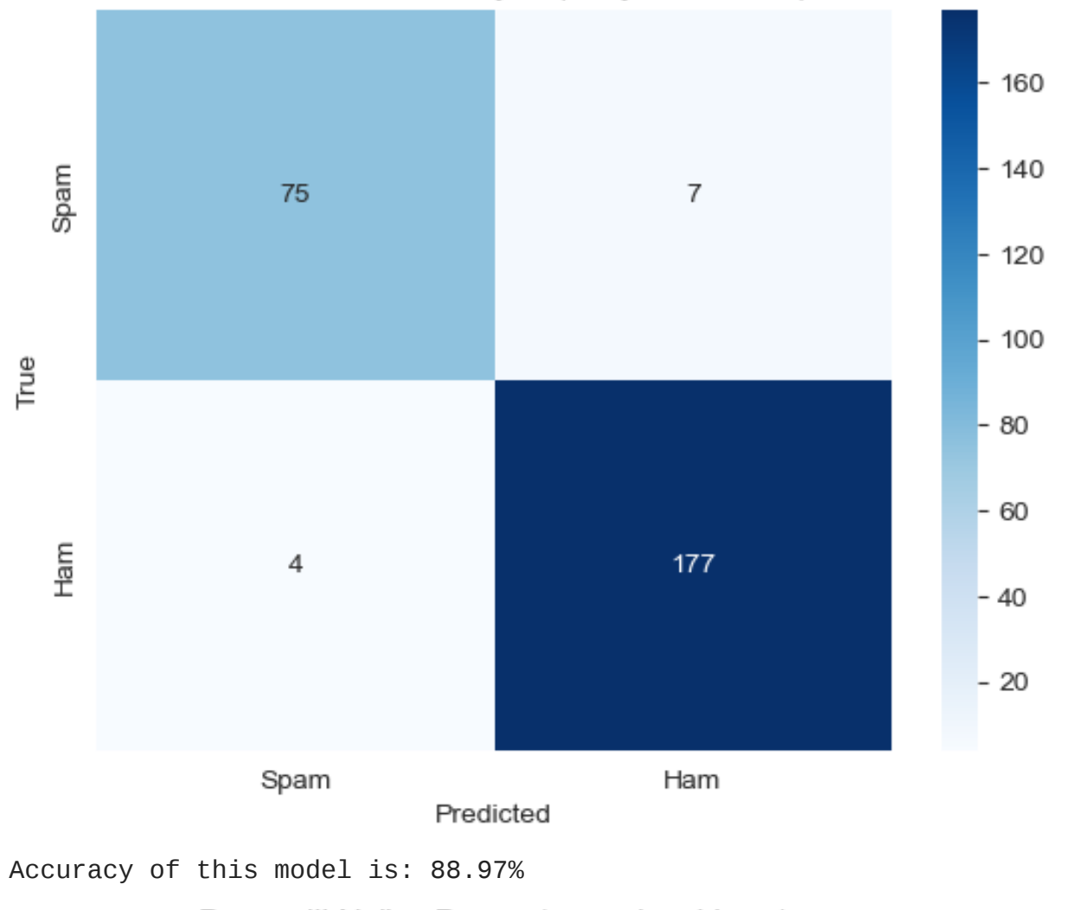
Yes. Both the models now have less accuracy than the one in question 2. This is because this model is trying to detect spam in hard ham which is harder to differentiate from spam.

3.2 Retrain

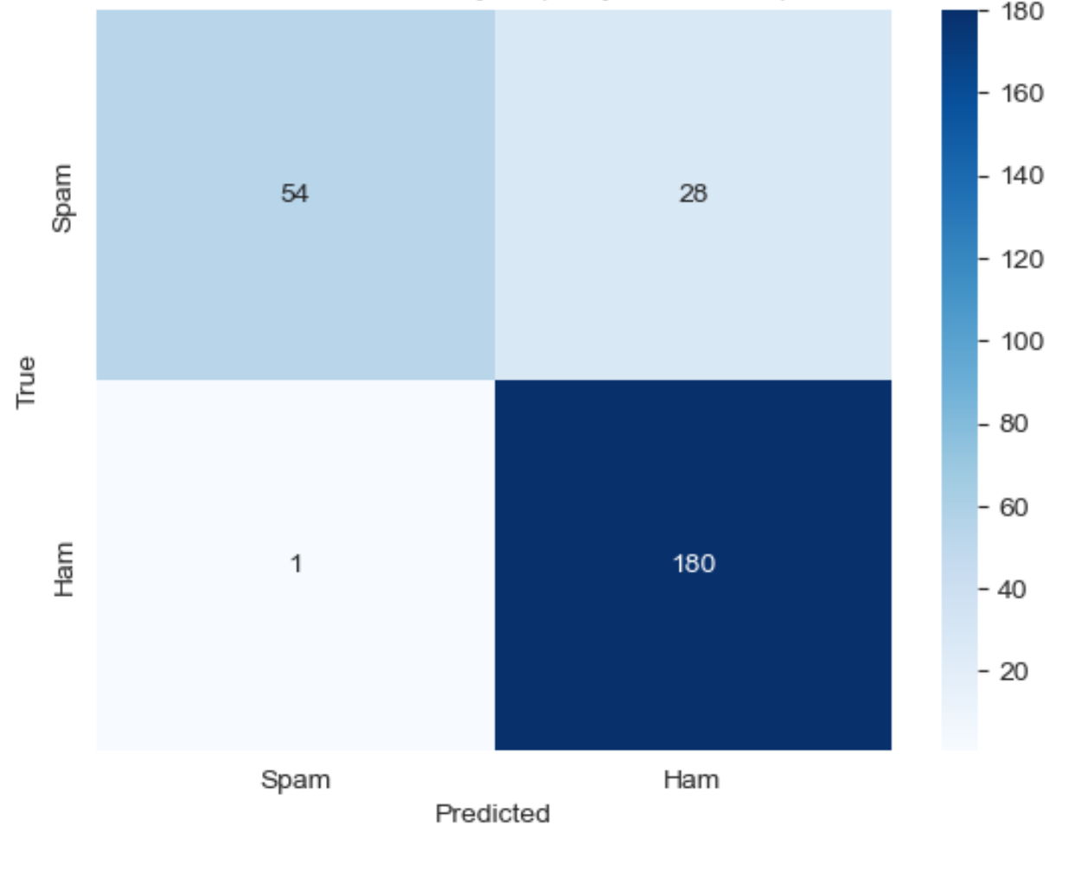
Retrain new Multinomial and Bernolli Naive Bayes classifers on the combined (easy+hard) ham and spam. Now evaluate on spam versus hard-ham as in 3.1. Also evaluate on spam versus easy-ham. Compare the performance with question 2 and 3.1. What do you observe?

In [14]:
```
df_h = extract_emails(['easy_ham'] and ['hard_ham'], 'ham')

df_ham_spam = pd.concat([df_h, df_s])

# spam = 1 and ham = 0
df_ham_spam.loc[df_ham_spam['class'] == 'spam', 'class'] = 1
df_ham_spam.loc[df_ham_spam['class'] == 'ham', 'class'] = 0

# Create datasets with 65& train size and 35% test size
hamtrain, hamtest, spamtrain, spamtest = train_test_split(df_ham_spam['content'], df_ham_spam['class'].values.tolist(), train_size=0.65)

spam_detection_cm(hamtrain, hamtest, spamtrain, spamtest, categories, group_names, "Multinomial Naïve Bayes (easy+hard ham)", 'MNB')
spam_detection_cm(hamtrain, hamtest, spamtrain, spamtest, categories, group_names, "Bernoulli Naïve Bayes (easy+hard ham)", 'BNB')
```

```
Accuracy of this model is: 95.82%
```



```
Accuracy of this model is: 88.97%
```



Answer 3.2:

The accuracy for the combined model becomes lower than for the one trained and tested on only easy ham but higher for the one tested with hard ham, for both classifiers. We are getting higher accuracy of the combined than the only hard ham because now we have a lot of mails that are easy ham, and the model is also trained by a larger dataset. The combined have less accuracy than the one trained and tested for easy ham because now we also have mails that are harder to detect as spam.

3.3 Further improvements

Answer 3.3:

The above models mainly rely on counting words. If we somehow inlcuded more parameters to the model such as detect the letter colors, font style (bold and cursive) and look after emojis since this is recurring in spam mails (by experience). 😅😜👾😈🥳😂